УДК 004.891.2

В. В. Уштаніт
А. А. Папа
А. А. Яровий
О.П. Прозор

АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ

Вінницький національний технічний університет

Анотація

Розглянуто актуальність проблеми прогнозування відтоку клієнтів. Здійснено аналіз методів машинного навчання для прогнозування відтоку клієнтів.

Ключові слова: машинне навчання, інтелектуальний аналіз даних, дерева рішень, kNN, ансамблі рішень.

Abstract

The urgency of the problem of forecasting outflow of clients is considered. The analysis of machine learning methods for predicting outflow of clients is carried out.

Keywords: machine learning, data mining, decision trees, kNN, decision ensemble.

Вступ

Натепер більшість продуктових компаній орієнтуються на отриманні нових та збереженні старих клієнтів, яким надаються ті чи інші послуги компанії. Однією із задач маркетингових відділів є прогнозування відтоку клієнтів.

Актуальність ідеї створення програмного продукту для прогнозування відтоку клієнтів є досить високо-пріоритетною на даний час, адже майже кожен власник бізнесу, який побудований на роботі з клієнтами та наданні певних послуг, хоче бути застрахований або, як мінімум, попереджений про можливий відтік клієнтів. Але досить часто доводиться зустрічатися з проблемою відсутності спеціалізованих програмних засобів для прогнозування відтоку клієнтів, що враховують специфіку предметної області.

Метою роботи є дослідження існуючих методів машинного навчання, а також покращення прогнозу відтоку клієнтів за допомогою використання методів машинного навчання.

Результати дослідження

Розглянемо детальніше саме поняття машинного навчання (МН). МН – це напрям штучного інтелекту, що розглядає побудову алгоритмів, які можуть навчатися на наявних даних [1]. Задача МН виглядає так: нехай є певний набір об’єктів – прикладів і певний набір міток, тобто, реакцій, відповідей. Між прикладами/спостереженнями і відповідями є певна прихована залежність. Задача МН – знайти цю приховану залежність для прогнозування відповідей на основі нових даних.

МН поділяється на 3 типи:

1) навчання з вчителем – є набір прикладів, до кожного прикладу є правильна відповідь. Задача системи – навчитися по прикладах надавати правильну відповідь, задану вчителем.

2) навчання без вчителя – є великий набір даних. В цих даних є приховані закономірності. Задача системи – знайти закономірності, наприклад, розбивши дані на певні групи чи кластери.

3) навчання з підкріпленням – програма взаємодіє з динамічним середовищем, у якому вона повинна виконувати певну мету без учителя, який явно казав би їй, чи підійшла вона близько до мети. Прикладом є навчання гри через гру із суперником.

Задачі МН класифікуються ще на декілька типів за видом вирішуваної проблеми [2]:

- задача класифікації – віднесення об’єкту до однієї з категорій на основі її характеристик;
- задача кластеризації – розбиття множини об’єктів на групи на основі характеристик цих об’єктів таким чином, щоб в одній групі були схожі між собою об’єкти, і менш схожі з об’єктами інших груп;
- задача регресії – прогнозування кількісної характеристики об’єкта на основі інших його характеристик;
- задача виявлення аномалій – пошук об’єктів, «сильно не схожих» на всі інші у вибірці або на якусь групу об’єктів.

Задача прогнозування відтоку клієнтів є задачею класифікації. Тобто, на основі відомих характеристик користувача необхідно передбачити належність його до групи тих користувачів, які підуть або залишаться. Задача класифікації є задачею навчання з вчителем, тобто необхідні наявні набори даних: навчальна та тестова вибірки. Найпопулярніші методи МН для вирішення задачі класифікації – це дерева рішень та метод найближчих сусідів.

Дерево рішень – це елементарний класифікатор, який є об’єднанням логічних правил типу «Якщо значення *а* менше *х* та *в* більше *у,* то *клас 1*» в структуру даних «дерево».

Даний метод є досить інтерпретуємим, тобто, є наочним. Також, дерева рішень можуть легко візуалізуватись та візуалізувати конкретне рішення в ньому. Процес навчання та прогнозування є досить швидким при невеликій кількості параметрів. Також, перевагою є підтримка числових та категоріальних показників.

Серед недоліків можна виділити чутливість до навчальних даних. Дерево може повністю змінитись, якщо навчальна вибірка трохи зміниться. Також, необхідно відсікати гілки дерева або встановлювати мінімальне число елементів в листках дерева або максимальну глибину дерева для боротьби з перенавчанням. Присутня, також, проблема пошуку оптимального дерева рішень (на практиці часто використовують евристики, типу жадібного пошуку, які не гарантують знаходження оптимального дерева).

Ще одним методом є метод найближчих сусідів (k Nearest Neighbors, або kNN) – відносно простий непараметричний класифікаційний метод, де для класифікації об'єктів у рамках простору властивостей використовуються відстані (зазвичай, евклідові), пораховані до усіх інших об'єктів. Вибираються об'єкти, до яких відстань найменша, і вони виділяються в окремий клас.

Метод k-найближчих сусідів – метричний алгоритм для автоматичної класифікації об'єктів. Основним принципом методу найближчих сусідів є те, що об'єкт присвоюється тому класу, який є найбільш поширеним серед сусідів даного елемента. Сусіди беруться, виходячи з множини об'єктів, класи яких уже відомі, і, виходячи з ключового для даного методу значення k, вираховується, який клас є найчисленнішим серед них. Кожен об'єкт має кінцеву кількість атрибутів (розмірностей). Передбачається, що існує певний набір об'єктів з уже наявною класифікацією.

До переваг можна виділити те, що метод є простим в реалізації, добре вивчений теоретично, є зручним методом для первинного вирішення задачі, його можна адаптувати під потрібну задачу вибором метрики або методу оцінки подібності, також метод є достатню інтерпретуємим.

До недоліків можна віднести відносно низьку швидкодію методу на великій кількості даних. При великій кількості параметрів об’єкту тяжко підібрати відповідні ваги для характеристик. Результат сильно залежить від вибору метрики відстані.

Наступною є група лінійних моделей, які використовують лінійну та логістичну регресію, методи найменших квадратів та максимальної правдоподібності [3].

Ці методи є достатньо добре вивченими, працюють дуже швидко та працюють на дуже великих вибірках. Досить ефективні, коли наявна велика кількість характеристик об’єктів. Також, логістична регресія видає вірогідності віднесення до різних класів.

Недоліками є те, що ця група методів працює недостатньо ефективно, коли присутня нелінійна залежність відповідей від характеристик об’єктів.

Останнім часом стала популярною техніка об’єднання декількох методів МН. Такі композиції називають ансамблями [4]. Суть полягає в розбитті вибірки на частини, які обробляються різними алгоритмами МН, а потім підсумковий класифікатор усереднює відповіді всіх алгоритмів.

Популярним видом ансамблів є випадковий ліс рішень. Полягає він у побудові певної кількості дерев рішень з різних шматків навчальних даних. Для того, щоб використати даний ансамбль,

використовуються всі дерева рішень та об’єднуються результати за допомогою мажоритарного голосування або комбінуванням апостеріорних ймовірностей.

Згідно описаного вище, зробимо висновок, що для прогнозування відтоку клієнтів, найдоцільніше використовувати метод навчання з вчителем, на основі вхідного та вихідного наборів даних, з використанням ансамблю з різних методів МН для конкретного розподілу користувачів на категорії.

Висновки

Здійснено аналіз актуальності задачі прогнозування відтоку клієнтів за допомогою методів машинного навчання. Досліджено та проаналізовано основні задачі та методи машинного навчання, а також їх переваги і недоліки. Визначено доцільність використання ансамблю рішень для задачі прогнозування відтоку клієнтів.

Отримані результати дослідження показують доцільність і перспективність застосування обраних методів для створення реального програмного продукту. Отримані результати планується використати в подальшій роботі з метою підвищення якості прогнозування відтоку клієнтів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Машинне навчання. Типи навчання. [Електронний ресурс]. – Режим доступу: https://courses.prometheus.org.ua/courses/IRF/ML101/2016_T3/about
2. MachineLearning.ru [Електронний ресурс]. – Режим доступу: http:// www.machinelearning.ru
3. Binder, D.A. Bayesian cluster analysis. - Biometrika, 1978. - 65, 31–38.
4. Jerome H. Friedman Elements of Statistical Learning [Електронний ресурс] / Robert Tibshirani // Режим доступу:URL: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Уштаніт Вадим Вікторович — студент групи 2КН-17м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: vadim.ushtanit@gmail.com.

Папа Андрій Андрійович — студент групи 2КН-17м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: papa.andriy@gmail.com.

Яровий Андрій Анатолійович — д.т.н., професор, завідувач кафедри комп’ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: a.yarovyy@vntu.edu.ua.

Прозор Олена Петрівна — к.пед.н., доцент, доцент кафедри комп’ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: prozor@vntu.edu.ua.

Vadim V. Ushtanit — student of Information Technologies and Computer Engineering Department, 2CS-17m, Vinnytsia National Technical University, Vinnytsia, e-mail: vadim.ushtanit@gmail.com.

Andrii A. Papa — student of Information Technologies and Computer Engineering Department, 2CS- 17m, Vinnytsia National Technical University, Vinnytsia, e-mail: papa.andriy@gmail.com.

Andrii A. Yarovyi — Doctor of Science (Eng.), Professor, Head of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail: a.yarovyy@vntu.edu.ua.

Olena P. Prozor — PhD (Eng.), Professor Assistant, Professor Assistant of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail: prozor@vntu.edu.ua.